CONVERGENCE OF GRADIENT-BASED ALGORITHMS FOR THE HARTREE-FOCK EQUATIONS

ANTOINE LEVITT

ABSTRACT. The numerical solution of the Hartree-Fock equations is a central problem in quantum chemistry for which numerous algorithms exist. Attempts to justify these algorithms mathematically have been made, notably in [5], but, to our knowledge, no complete convergence proof has been published. In this paper, we prove the convergence of a natural gradient algorithm, using a gradient inequality for analytic functionals due to Łojasiewicz [15]. Then, expanding upon the analysis of [5], we prove convergence results for the Roothaan and Level-Shifting algorithms. In each case, our method of proof provides estimates on the convergence rate. We compare these with numerical results for the algorithms studied.

1. INTRODUCTION

In quantum chemistry, the Hartree-Fock method is one of the simplest approximations of the electronic structure of a molecule. By assuming minimal correlation between the N electrons, it reduces Schrödinger's equation, a linear partial differential equation on $\mathbb{R}^{3N}$, to the Hartree-Fock equations, a system of N coupled nonlinear equations on $\mathbb{R}^3$. This approximation makes it much more tractable numerically. It is used both as a standalone description of the molecule and as a starting point for more advanced methods, such as the Møller-Plesset perturbation theory, or multi-configuration methods. Mathematically, the Hartree-Fock method leads to a coupled system of nonlinear integro-differential equations, which are discretized by expanding the solution on a finite Galerkin basis. The resulting nonlinear algebraic equations are then solved iteratively, using a variety of algorithms, the convergence of which is the subject of this work.

The mathematical structure of the Hartree-Fock equations was investigated in the 70's, culminating in the proof of the existence of solutions by Lieb and Simon [13], later generalized by Lions [14]. On the other hand, despite their ubiquitous use in computational chemistry, the convergence of the various algorithms used to solve them is still poorly understood. A major step forward in this direction is the recent work of Cancès and Le Bris [5]. Using the density matrix formulation, they provided a mathematical explanation for the oscillatory behavior observed in the simplest algorithm, the Roothaan method, and proposed the Optimal Damping Algorithm (ODA), a new algorithm inspired directly by the mathematical structure of the constraint set [4]. This algorithm was designed to decrease the energy at each step, and linking the energy decrease to the difference of iterates allowed the authors to prove that this algorithm "numerically converges" in the weak sense that $\|D_k - D_{k-1}\| \to 0$. In addition, the algorithm numerically converges towards an Aufbau solution [7]. This, to our knowledge, is the strongest convergence result available for an algorithm to solve the Hartree-Fock equations.

However, this is still mathematically unsatisfactory, as it does not guarantee convergence, and merely prohibits fast divergence. The difficulty in proving convergence of the algorithms used to solve the Hartree-Fock equations lies in the lack of understanding of the second-order properties of the energy functional (for instance, there are no local uniqueness results available). In other domains, the convergence of gradient-based methods has been established using the Łojasiewicz inequality for analytic functionals [15] (see for instance [10, 18]). This method of proof has the advantage of not requiring any second-order information.

In this paper, we use a gradient descent algorithm to solve the Hartree-Fock equations. This algorithm builds upon ideas from differential geometry [8] and the various projected gradient algorithms used in the context of quantum chemistry [1, 6, 16]. To our knowledge, this particular algorithm has never been applied to the Hartree-Fock equations. Although it lacks the sophistication of modern minimization

methods (for instance, see the trust region methods of [9] and [11]), it is the most natural generalization of the classical gradient descent, and, as such, the simplest one to analyze mathematically. For this algorithm, following the method of [18], we prove convergence, and obtain explicit estimates on the convergence rate. We also apply the method to the widely used Roothaan and Level-Shifting algorithms, effectively linking these fixed-point algorithms to gradient methods.

The structure of this paper is as follows. We first introduce the Hartree-Fock problem in the mathematical setting of density matrices and prove a Łojasiewicz inequality on the constrained parameter space. We then introduce the gradient algorithm, and prove some estimates. We show the convergence and obtain convergence rates for this algorithm, then extend our method to the Roothaan and Level-Shifting algorithm, using an auxiliary energy functional following [5]. We finally test all these results numerically and compare the convergence of the algorithms.

2. Setting

We are concerned with the numerical solution of the Hartree-Fock equations. We will consider for simplicity of notation the spinless Hartree-Fock equations, where each orbital ϕ_i is a function in $L^2(\mathbb{R}^3, \mathbb{R})$, although our results are easily transposed to other variants such as General Hartree-Fock (GHF) and Restricted Hartree-Fock (RHF).

In this paper, we consider a Galerkin discretization space with finite orthonormal basis $(\chi_i)_{i=1\ldots N_b}$. In this setting, the orbitals ϕ_i are expanded on this basis, and the operators we consider are $N_b \times N_b$ matrices. This finite dimension hypothesis is crucial for our results, and we comment on it in the conclusion.

The Hartree-Fock problem consists in minimizing the total energy of a N-body system. We describe the mathematical structure of the energy functional and the minimization set, and introduce a natural gradient descent to solve this problem numerically.

2.1. **The energy.** We consider the quantum N-body problem of N electrons in a potential V (in most applications, V is the Coulombic potential created by a molecule or atom). In the spinless Hartree-Fock model, this problem is simplified by assuming that the N-body wavefunction ψ is a single Slater determinant of N L^2-orthonormal orbitals ϕ_i. A simple calculation then shows that the energy of the wavefunction ψ can be expressed as a function of the orbitals ϕ_i,

$$\mathcal{E}(\psi) = \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{1}{2}(\nabla \phi_i)^2 + \int_{\mathbb{R}^3} V\rho + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y) - \tau(x,y)^2}{|x-y|} \, \mathrm{d}x \, \mathrm{d}y,$$

where $\tau(x,y) = \sum_{i=1}^{N} \phi_i(x)\phi_i(y)$ and $\rho(x) = \tau(x,x)$.

The energy is then to be minimized over all sets of orthonormal orbitals ϕ_i. An alternative way of looking at this problem is to reformulate it using the density operator D. This operator, defined by its integral kernel $D(x,y) = \tau(x,y)$, can be seen to be the orthogonal projection on the space spanned by the ϕ_i's. The energy can be written as a function of D only:

$$E(D) = \mathrm{Tr}((h + \frac{1}{2}G(D))D), \tag{2.1}$$

where

$$h = -\frac{1}{2}\Delta + V,$$

$$(G(D)\phi)(x) = \left(\rho \star \frac{1}{|\cdot|}\right)(x)\phi(x) - \int_y \frac{\phi(y)\tau(x,y)}{|x-y|}.$$

This time, the energy is to be minimized over all orthogonal projection operators of rank N. In the discrete setting, the orbitals ϕ_j are discretized as $\phi_j = \sum_{i=1}^{N_b} c_{ij}\chi_i$, and the operators D, h, and $G(D)$ become $N_b \times N_b$ matrices.

2.2. **The manifold $\mathcal{P}$.** The Hartree-Fock energy is to be minimized over the set of density matrices

$$D \in \mathcal{P} = \{D \in M_{N_b}(\mathbb{R}), D^T = D, D^2 = D, \mathrm{Tr}\, D = N\}.$$

The manifold $\mathcal{P}$ is equipped with the canonical inner product in $M_{N_b}(\mathbb{R})$

$$\langle A, B \rangle = \mathrm{Tr}(A^T B).$$

We denote by $\|A\| = \sqrt{\langle A, A \rangle}$ the Frobenius (or Hilbert-Schmidt) norm of A, which is the most natural here, and by $\|A\|_{\mathrm{op}} = \max_{\|x\|=1} \|Ax\|$ the operator norm of A.

The Riemannian structure of $\mathcal{P}$ allows us to compute the gradient of E. The tangent space $T_D\mathcal{P}$ at a point D is the set of Δ such that $D + \Delta$ verifies the constraints up to first order in Δ, that is, such that $\Delta^T = \Delta, D\Delta + \Delta D = \Delta, \operatorname{Tr}\Delta = 0$. Block-decomposing Δ on the two orthogonal spaces range(D) and ker(D), this implies that the tangent space $T_D\mathcal{P}$ is the set of matrices Δ of the form

$$\Delta = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix},$$

where A is an arbitrary $(N_b - N) \times N$ matrix.

Hence, the projection on the tangent space of an arbitrary symmetric matrix M is given by

$$P_D(M) = DM(1 - D) + (1 - D)MD$$
$$= [D, [D, M]].$$

Note that if M has decomposition $\begin{pmatrix} B & A^T \\ A & C \end{pmatrix}$, then $[D, M] = \begin{pmatrix} 0 & A^T \\ -A & 0 \end{pmatrix}$ and $[D, [D, M]] = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$, which shows that $\big\|[D[D, M]]\big\| = \big\|[D, M]\big\|$, a property that will be useful in the sequel.

We can now compute the gradient of E. First, the unconstrained gradient in $M_{N_b}(\mathbb{R})$ is

$$\nabla E(D) = F_D = h + G(D),$$

the Fock operator describing the mean field generated by the electrons of D. We obtain the constrained gradient $\nabla_{\mathcal{P}} E$ by projecting onto the tangent space:

$$\nabla_{\mathcal{P}} E(D) = P_D(\nabla E(D))$$
$$= [D, [D, F_D]].$$

2.3. **Łojasiewicz inequality.** The Łojasiewicz inequality for a functional f around a critical point x_0 is a local inequality that provides a lower bound on ∇f. Its only hypothesis is analyticity. In particular, no second order information is needed, and the inequality accommodates degenerate critical points.

2.3.1. *The classical Łojasiewicz inequality.*

Theorem 2.1 (Łojasiewicz inequality in $\mathbb{R}^n$). *Let f be an analytic functional from $\mathbb{R}^n$ to $\mathbb{R}$. Then, for each $x_0 \in \mathbb{R}^n$, there is a neighborhood U of x_0 and two constants $\kappa > 0$, $\theta \in (0, 1/2]$ such that when $x \in U$,*

$$\big|f(x) - f(x_0)\big|^{1-\theta} \leq \kappa\big\|\nabla f(x)\big\|.$$

This inequality is trivial when x_0 is not a critical point. When x_0 is a critical point, a simple Taylor expansion shows that, if the Hessian $\nabla^2 f(x_0)$ is invertible, we can choose $\theta = \frac{1}{2}$ and $\kappa > \frac{1}{\sqrt{2|\lambda_1|}}$, where λ_1 is the eigenvalue of lowest magnitude $\nabla^2 f(x_0)$. When $\nabla^2 f(x_0)$ is singular (meaning that x_0 is a degenerate critical point), the analyticity hypothesis ensures that the derivatives cannot all vanish simultaneously, and that there exists a differentiation order n such that the inequality holds with $\theta = \frac{1}{n}$.

2.3.2. *Łojasiewicz inequality on $\mathcal{P}$.* We now extend this inequality to functionals defined on the manifold $\mathcal{P}$.

Theorem 2.2 (Łojasiewicz inequality on $\mathcal{P}$). *Let f be an analytic functional from $\mathcal{P}$ to $\mathbb{R}$. Then, for each $D_0 \in \mathcal{P}$, there is a neighborhood U of D_0 and two constants $\kappa > 0$, $\theta \in (0, 1/2]$ such that when $D \in U$,*

$$\big|f(D) - f(D_0)\big|^{1-\theta} \leq \kappa\big\|\nabla_{\mathcal{P}} f(D)\big\|.$$

Proof. Let $D_0 \in \mathcal{P}$. Define the map R_{D_0} from $T_{D_0}\mathcal{P}$ to $\mathcal{P}$ by

$$R_{D_0}(\Delta) = UD_0U^T,$$
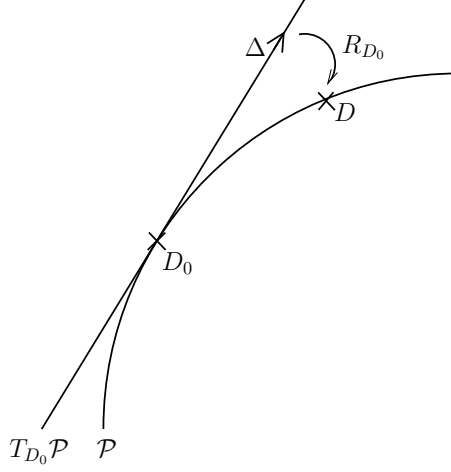$$U = \exp(-[D_0, \Delta]).$$

FIGURE 2.1. The map R_{D_0}

This map is analytic and verifies $R_{D_0}(0) = D_0$, $\mathrm{d}R_{D_0}(0) = \mathrm{id}_{T_{D_0}\mathcal{P}}$. Therefore, by the inverse function theorem, the image of a neighborhood of zero contains a neighborhood of D_0. We now compute the gradient of $\widetilde{f} = f \circ R_{D_0}$ at a point Δ, with $D = R_{D_0}(\Delta)$.

$$\widetilde{f}(\Delta + \delta) = f(D) + \langle \nabla_{\mathcal{P}} f(D), \mathrm{d}R_{D_0}(\Delta)\delta \rangle + O(\delta^2)$$
$$= f(D) + \langle \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D), \delta \rangle + O(\delta^2)$$
$$= f(D) + \langle P_{D_0} \, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D), \delta \rangle + O(\delta^2).$$

We deduce

$$\nabla_{T_{D_0}\mathcal{P}} \widetilde{f}(\Delta) = P_{D_0} \, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D).$$

We can now apply the Łojasiewicz inequality to $\widetilde{f}$, which is an analytic functional defined on the Euclidean space $T_{D_0}\mathcal{M}$. We obtain a neighborhood of zero in $T_{D_0}\mathcal{P}$, and therefore a neighborhood U of D_0 on which

$$\left| f(D) - f(D_0) \right|^{1-\theta} \leq \kappa \left\| P_{D_0} \, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D) \right\|.$$

As $\mathrm{d}R_{D_0}^*$ is continuous in Δ, up to a change in the neighborhood U and the constant κ,

$$\left| f(D) - f(D_0) \right|^{1-\theta} \leq \kappa \left\| \nabla_{\mathcal{P}} f(D) \right\|.$$

$\square$

3. THE GRADIENT METHOD

3.1. **Description of the method.** The gradient flow

$$\frac{\mathrm{d}D}{\mathrm{d}t} = -\nabla_{\mathcal{P}} E(D)$$
$$= -[D, [D, F_D]] \tag{3.1}$$

is a way of minimizing the energy over the manifold $\mathcal{P}$. This continuous flow was already used to solve the Hartree-Fock equations in [1] (although the authors used a formulation in terms of orbitals, whereas we use density matrices).

The naive Euler discretization

$$D_{k+1} = D_k - t[D_k, [D_k, F_k]]$$

is not suitable because it does not stay on $\mathcal{P}$. A variety of approaches deal with this problem. One of the first algorithms to solve the Hartree-Fock equations [16] used a purification method to project D_{k+1} back onto $\mathcal{P}$. More recently, an orthogonal projection on the convex hull of $\mathcal{P}$ was used for that purpose [6]. Although we focus in this paper on a different gradient method, such projection methods have the same behavior for small stepsizes and can be treated in the same framework, provided one can prove results similar to Lemmas 3.1 and 3.2 below.

4

We look for D_{k+1} on a curve on $\mathcal{P}$ that is tangent to $\nabla_{\mathcal{P}} E(D_k)$. A natural curve on $\mathcal{P}$ is the change of basis

$$D'(t) = U_t D U_t^T,$$

where U_t is a smooth family of orthogonal matrices. If we take

$$U_t = \exp(t[D, F_D]),$$

we get

$$\left. \frac{\mathrm{d} D'}{\mathrm{d} t} \right|_{t=0} = -[D, [D, F_D]],$$

so $D'(t)$ is a smooth curve on $\mathcal{P}$, tangent to the gradient flow at $t = 0$.

Our gradient method with a fixed step t is then

$$D_{k+1} = U_k D_k U_k^T, \tag{3.2}$$

with

$$U_k = \exp(t[D_k, F_k]). \tag{3.3}$$

This method, as well as various generalizations, is described in [8].

We now prove a number of lemmas which are the main ingredients of the convergence proof. First, we bound the second derivative of the energy to obtain quantitative estimates on the energy decrease, then we link the difference of iterates $D_{k+1} - D_k$ to the gradient $\nabla_{\mathcal{P}} E(D_k)$, and finally we use the Łojasiewicz inequality to establish convergence.

3.2. **Derivatives.** We start from a point D_0 and compute the derivatives of E along the curve $D_t = U_t D_0 U_{-t}$. For ease of notation we will write $\epsilon(t) = E(D_t)$, $F_t = F(D_t)$ and $C_t = [D_t, F_t]$.

$$
\begin{aligned}
\frac{\mathrm{d} D_t}{\mathrm{d} t} &= \frac{\mathrm{d} U_t}{\mathrm{d} t} D_0 U_{-t} + U_t D_0 \frac{\mathrm{d} U_{-t}}{\mathrm{d} t} \\
&= [C_0, D_t],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}^n D_t}{\mathrm{d} t^n} &= \frac{\mathrm{d}^{n-1}}{\mathrm{d} t^{n-1}} [C_0, D_t] \\
&= \underbrace{[C_0, [C_0, \dots [C_0, D_t] \dots]]}_{n \text{ times } C_0},
\end{aligned}
$$

$$
\frac{\mathrm{d} \epsilon}{\mathrm{d} t} = \mathrm{Tr}(F_t[C_0, D_t]),
$$

$$
\left. \frac{\mathrm{d} \epsilon}{\mathrm{d} t} \right|_{t=0} = -\|C_0\|^2,
$$

$$
\frac{\mathrm{d}^2 \epsilon}{\mathrm{d} t^2} = \mathrm{Tr}(F_t[C_0, [C_0, D_t]]) + \mathrm{Tr}(G([C_0, D_t])[C_0, D_t]).
$$

3.3. **Control on the curvature.**

Lemma 3.1. *There exists $\alpha > 0$ such that for every D_0 and t,*

$$\left| \frac{\mathrm{d}^2 \epsilon}{\mathrm{d} t^2} \right| (t) \leq \alpha \|C_0\|^2.$$

Proof.

$$\frac{\mathrm{d}^2 \epsilon}{\mathrm{d} t^2} = \mathrm{Tr}(F_t[C_0, [C_0, D_t]]) + \mathrm{Tr}(G([C_0, D_t])[C_0, D_t]). \tag{3.4}$$

First, since the Laplacian in $F(D)$ acts on a finite dimensional space, we can bound $F(D)$:

$$
\begin{aligned}
\left\| F(D) \right\|_{\mathrm{op}} &\leq \frac{1}{2} \|-\Delta\|_{\mathrm{op}} + \|V\|_{\mathrm{op}} + \|G(D)\|_{\mathrm{op}} \\
&\leq \frac{1}{2} \|-\Delta\|_{\mathrm{op}} + 2(2N + Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}
\end{aligned} \tag{3.5}
$$

by the Hardy inequality. Next, making use of the operator inequality $\mathrm{Tr}(AB) \leq \|A\|_{\mathrm{op}}\|B\|$, we show that

$$\mathrm{Tr}(F_t[C_0, [C_0, D_t]]) \leq 2\left(\frac{1}{2}\|-\Delta\|_{\mathrm{op}} + 2(2N + Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}\right)\|C_0\|^2.$$

For the second term of (3.4),

$$\begin{aligned}
\mathrm{Tr}(G([C_0, D_t])[C_0, D_t]) &\leq \|G([C_0, D_t])\|_{\mathrm{op}}\mathrm{Tr}\left(|[C_0, D_t]|\right)\\
&\leq 4\sqrt{\|-\Delta\|_{\mathrm{op}}}\,\mathrm{Tr}\left(|[C_0, D_t]|\right)^2\\
&\leq 16N\sqrt{\|-\Delta\|_{\mathrm{op}}}\|C_0\|^2.
\end{aligned}$$

The result is now proved with

$$\alpha = \|-\Delta\|_{\mathrm{op}} + 4(6N + Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}.$$

$\square$

3.4. **Choice of the stepsize.** We can now expand the energy:

$$\epsilon(t) \leq \epsilon(0) - t\|C_0\|^2 + \frac{t^2}{2}\alpha\|C_0\|^2.$$

If we choose

$$t < \frac{2}{\alpha}, \tag{3.6}$$

we obtain a decrease of the energy

$$\epsilon(t) \leq \epsilon(0) - \beta\|C_0\|^2 \tag{3.7}$$

with $\beta = t - \frac{t^2}{2}\alpha > 0$.

The optimal choice of t with this bound on the curvature is $t = \frac{1}{\alpha}$, with $\beta = \frac{1}{2\alpha}$. Of course it could be that the actual optimal t is different, and could vary wildly, which is why we will not consider optimal stepsizes.

3.5. **Study of $D_{k+1} - D_k$.** We now prove that our iteration $D_{k+1} = U_k D_k U_k^T$ coincides with an unconstrained gradient method up to first order in t.

We say that $M \in o(\|N\|)$ when for all $\varepsilon > 0$, there is a neighborhood U of zero such that when $N \in U$, $\|M\| \leq \varepsilon\|N\|$. Note that this neighborhood U is not allowed to depend on N, meaning that the resulting bound is uniform, which will allow us to manipulate the remainders more easily.

Lemma 3.2. *For any k,*

$$D_{k+1} = D_k + t[C_k, D_k] + o(t\|C_k\|).$$

Proof.

$$D_{k+1} - D_k - t[C_k, D_k] = \sum_{n=2}^{\infty}\frac{t^n}{n!}\underbrace{[C_k, [C_k, \dots [C_k, D_k]\dots]]}_{n \text{ times } C_k}$$

$$\left\|D_{k+1} - D_k - t[C_k, D_k]\right\| \leq t\big\|[C_k, D_k]\big\|\sum_{n=2}^{\infty}t^{n-1}\|C_k\|^{n-1}$$

$$\leq t\big\|[C_k, D_k]\big\|\frac{t\|C_k\|}{1 - t\|C_k\|}$$

$\square$

4. Convergence of the gradient algorithm

Theorem 4.1 (Convergence of the gradient algorithm). *Let $D_0 \in \mathcal{P}$ be any density matrix and D_k be the sequences of iterates generated from D_0 by $D_{k+1} = U_k D_k U_k^T$, with stepsize $t < \frac{2}{\alpha}$. Then D_k converges towards a solution of the Hartree-Fock equations.*

Proof. The energy $E(D)$ is bounded from below on $\mathcal{P}$, and therefore E_k converges to a limit E_∞. In the sequel we will work for convenience with $\widetilde{E}(D) = E(D) - E_\infty$ and drop the tildes. Immediately, summing (3.7) implies that C_k converges to 0, and therefore so does $D_k - D_{k-1}$ (this is what Cancès and Le Bris call "numerical convergence" in [5]). Note that we only get that $\|D_k - D_{k-1}\|^2$ is summable, which alone is not enough to guarantee convergence (the harmonic series $x_k = \sum_{l=1}^{k} 1/l$ is a simple counter-example). To obtain convergence, we shall use the Łojasiewicz inequality.

Let us denote by Γ the level set $\Gamma = \{D \in \mathcal{P}, E(D) = 0\}$. It is non-empty and compact. We apply the Łojasiewicz inequality to every point of Γ to obtain a cover $(U_i)_{i \in \mathcal{I}}$ of Γ in which the Łojasiewicz inequality holds with constants κ_i, θ_i.

By compactness, we extract a finite subcover from the U_i, from which we deduce $\delta > 0$, $\kappa > 0$ and $\theta \in (0, 1/2]$ such that whenever $d(D, \Gamma) < \delta$,

$$E(D)^{1-\theta} \leq \kappa \|\nabla_\mathcal{P} E(D)\| = \kappa \|[D, C_D]\| = \kappa \|C_D\|. \tag{4.1}$$

(recall from Section 2.2 that $\|[D, [D, M]]\| = \|[D, M]\|$ for M symmetric)

To apply the Łojasiewicz inequality to our iteration, it remains to show that $d(D_k, \Gamma)$ tends to zero. Suppose this is not the case. Then we can extract a subsequence, still denoted by D_k, such that $d(D_k, \Gamma) \geq \varepsilon$ for some $\varepsilon > 0$. By compactness of $\mathcal{P}$ we extract a subsequence that converges to a D such that $d(D, \Gamma) \geq \varepsilon$ and (by continuity) $E(D) = 0$, a contradiction. Therefore $d(D_k, \Gamma) \to 0$, and for k larger than some k_0,

$$E(D_k)^{1-\theta} \leq \kappa \|C_k\|. \tag{4.2}$$

For $k \geq k_0$,

$$\begin{aligned}
E(D_k)^\theta - E(D_{k+1})^\theta &\geq \frac{\theta}{E(D_k)^{1-\theta}}(E(D_k) - E(D_{k+1})) \\
&\geq \frac{\theta}{\kappa \|C_k\|}(E(D_k) - E(D_{k+1})) \\
&\geq \frac{\theta \beta}{\kappa} \|C_k\| \\
&\geq \frac{\theta \beta}{\kappa t} \|D_{k+1} - D_k\| + o(\|D_{k+1} - D_k\|)
\end{aligned}$$

hence

$$\frac{\theta \beta}{\kappa t} \|D_{k+1} - D_k\| + o(\|D_{k+1} - D_k\|) \leq E(D_k)^\theta - E(D_{k+1})^\theta. \tag{4.3}$$

As the right-hand side is summable, so is the left-hand side, which implies that $\sum \|D_{k+1} - D_k\| < \infty$. D_k is therefore Cauchy and converges to some limit D_∞. $C_k \to 0$, so D_∞ is a critical point.

Note that now that we know that D_k converges to D_∞, we can replace the θ and κ in this inequality by the ones obtained from the Łojasiewicz inequality around D_∞ only. $\square$

Let

$$e_k = \sum_{l=k}^{\infty} \|D_{l+1} - D_l\|.$$

This is a (crude) measure of the error at iteration number k. In particular, $\|D_k - D_\infty\| \leq e_k$.

Theorem 4.2 (Convergence rate of the gradient algorithm).

(1) *If $\theta = 1/2$ (non-degenerate case), then for any $\nu' < \frac{\beta}{2\kappa^2}$, there exists $c > 0$ such that*

$$e_k \leq c(1 - \nu')^k. \tag{4.4}$$

(2) *If $\theta \neq 1/2$ (degenerate case), then there exists $c > 0$ such that*

$$e_k \leq c k^{-\frac{\theta}{1-2\theta}}. \tag{4.5}$$

7

Proof. Summing (4.3) from $l = k$ to ∞ with $k \geq k_0$, we obtain

$$e_k + o(e_k) \leq \frac{\kappa t}{\theta \beta} E(D_k)^\theta$$

$$\left(\frac{\theta \beta}{\kappa t} e_k + o(e_k) \right)^{\frac{1-\theta}{\theta}} \leq E(D_k)^{1-\theta}$$

$$\leq \kappa \|C_k\|$$

$$\leq \frac{\kappa}{t}(e_k - e_{k+1}) + o(e_k - e_{k+1})$$

Hence,

$$e_{k+1} \leq e_k - \nu e_k^{\frac{1-\theta}{\theta}} + o(e_k^{\frac{1-\theta}{\theta}}), \text{ with}$$

$$\nu = \frac{t}{\kappa} \left(\frac{\theta \beta}{\kappa t} \right)^{\frac{1-\theta}{\theta}}$$

Two cases are to be distinguished. If $\theta = \frac{1}{2}$, the above inequality reduces to

$$e_{k+1} \leq (1 - \nu + o(1))e_k$$

with $\nu = \frac{\beta}{2\kappa^2}$ and the result follows.

The case $\theta \neq 1/2$ is more involved. We define

$$y_k = ck^{-p},$$

which verifies

$$y_{k+1} = c(k+1)^{-p}$$

$$= ck^{-p}(1 + 1/k)^{-p}$$

$$\geq ck^{-p}(1 - \frac{p}{k})$$

$$\geq y_k(1 - pc^{-\frac{1}{p}} y_k^{\frac{1}{p}})$$

We set $p = \frac{\theta}{1-2\theta}$ and c large enough so that $c > (\frac{\nu}{p})^{-p}$ and $y_{k_0} \geq e_{k_0}$. We then prove by induction $e_k \leq y_k$ for $k \geq k_0$. The result follows by increasing c to ensure that $e_k \leq y_k$, for $k < k_0$. □

In the non-degenerate case $\theta = 1/2$ (which was the case for the numerical simulations we performed, see Section 7), the convergence is asymptotically geometric with rate $1 - \nu$, where

$$\nu = \frac{\beta}{2\kappa^2}.$$

With the choice $t = \frac{1}{\alpha}$ suggested by our bounds, the convergence rate is

$$\nu = \frac{1}{4\kappa^2 \alpha}.$$

5. Convergence of the Roothaan algorithm

The Roothaan algorithm (also called simple SCF in the literature) is based on the observation that a minimizer D of the energy satisfies the *Aufbau* principle: D is the projector onto the space spanned by the eigenvectors associated with the first N eigenvalues of $F(D)$. This suggests a simple fixed-point algorithm: take for D_{k+1} the projector onto the space spanned by the eigenvectors associated with the first N eigenvalues of $F(D_k)$, and iterate. Unfortunately, this procedure does not always work: in some circumstances, oscillations between two states occur, and the algorithm never converges. This behavior was explained mathematically in [5], where the authors notice that the Roothaan algorithm minimizes the bilinear functional

$$E(D, D') = \text{Tr}(h(D + D')) + \text{Tr}(G(D)D')$$

with respect to its first and second argument alternatively. Thanks to the Łojasiewicz inequality, we can improve on their result and prove the convergence or oscillation of the method.

The bilinear functional verifies $E(D, D') = E(D', D)$, $E(D, D) = 2E(D)$. In fact, $\frac{1}{2}E(\cdot, \cdot)$ is the symmetric bilinear form associated with the quadratic form $E(\cdot)$. In the following, we assume the uniform well-posedness hypothesis of [5], *i.e.* that there is a uniform gap of at least $\gamma > 0$ between the

eigenvalues number N and $N + 1$ of $F(D_k)$. Under this assumption, it can be proven [3] that there is a decrease of the bilinear functional at each iteration

$$
\begin{aligned}
E(D_{k+1}, D_{k+2}) &= E(D_{k+2}, D_{k+1}) \\
&= \min_{D \in \mathcal{P}} E(D, D_{k+1}) \\
&\leq E(D_k, D_{k+1}) - \gamma \|D_{k+2} - D_k\|^2
\end{aligned}
$$

Since $E(\cdot, \cdot)$ is bounded from below on $\mathcal{P} \times \mathcal{P}$, this immediately shows that $D_k - D_{k+2} \to 0$, which shows that D_{2k} and D_{2k+1} converge up to extraction, which was noted in [5]. We now prove convergence of these two subsequences, again using the Łojasiewicz inequality.

$E(\cdot, \cdot)$ is defined on $\mathcal{P} \times \mathcal{P}$, which inherits the Riemannian structure of $\mathcal{P}$ by the natural inner product $\langle (D_1, D_1'), (D_2, D_2') \rangle = \langle D_1, D_2 \rangle + \langle D_1', D_2' \rangle$. In this setting, the gradient is

$$
\nabla_{\mathcal{P} \times \mathcal{P}} E(D, D') = \begin{pmatrix} [D, F(D')] \\ [D', F(D)] \end{pmatrix}.
$$

and therefore, using the fact that D_{k+1} (resp. D_{k+2}) and $F(D_k)$ (resp $F(D_{k+1})$) commute,

$$
\begin{aligned}
\left\| \nabla_{\mathcal{P} \times \mathcal{P}} E(D_k, D_{k+1}) \right\| &= \sqrt{\left\| [D_k, F(D_{k+1})] \right\|^2 + \left\| [D_{k+1}, F(D_k)] \right\|^2} \\
&= \left\| [D_k, F(D_{k+1})] \right\| \\
&= \left\| [D_k - D_{k+2}, F(D_{k+1})] \right\| \\
&\leq 2 \left\| F(D_{k+1}) \right\|_{\mathrm{op}} \|D_{k+2} - D_k\|
\end{aligned}
$$

A trivial extension of Theorem 2.2 to the case of a functional defined on $\mathcal{P} \times \mathcal{P}$ shows that we can apply the Łojasiewicz inequality to $E(\cdot, \cdot)$. By the same compactness argument as before, the inequality

$$
\begin{aligned}
E(D_k, D_{k+1})^{1-\theta'} &\leq \kappa' \left\| \nabla_{\mathcal{P} \times \mathcal{P}} E(D_k, D_{k+1}) \right\| \\
&\leq 2\kappa' \left\| F(D_{k+1}) \right\|_{\mathrm{op}} \|D_{k+2} - D_k\|
\end{aligned}
$$

holds for k large enough, with constants $\kappa' > 0$ and $\theta' \in (0, \frac{1}{2}]$.

The exact same reasoning as for the gradient algorithm proves the following theorems

Theorem 5.1 (Convergence/oscillation of the Roothaan algorithm). *Let $D_0 \in \mathcal{P}$ such that the sequence D_k of iterates generated by the Roothaan algorithms verifies the uniform well-posedness hypothesis with uniform gap $\gamma > 0$. Then the two subsequences D_{2k} and D_{2k+1} are convergent. If both have the same limit, then this limit is a solution of the Hartree-Fock equations.*

Theorem 5.2 (Convergence rate of the Roothaan algorithm). *Let D_k be the sequence of iterates generated by a uniformly well-posed Roothaan algorithm, and let*

$$
e_k = \sum_{l=k}^{\infty} \|D_{l+2} - D_l\|.
$$

Then,

(1) If $\theta' = 1/2$ (non-degenerate case), then for any $\nu' < \frac{\gamma}{8\kappa'^2 \|F\|_{op}^2}$, where $\|F\|_{op}$ is the uniform bound on F proved in (3.5), there exists $c > 0$ such that

$$
e_k \leq c(1 - \nu')^k. \tag{5.1}
$$

(2) If $\theta' \neq 1/2$ (degenerate case), then there exists $c > 0$ such that

$$
e_k \leq c k^{-\frac{\theta'}{1-2\theta'}}. \tag{5.2}
$$

6. Level-shifting

The Level-Shifting algorithm was introduced in [19] as a way to avoid oscillation in self-consistent iterations. By shifting the N lowest energy levels (eigenvalues of F), one can force convergence, although denaturing the equations in the process. We now prove the convergence of this algorithm.

The same arguments as before apply to the functional

$$E^b(D, D') = \operatorname{Tr}(h(D + D')) + \operatorname{Tr}(G(D)D') + \frac{b}{2}\|D - D'\|^2$$
$$= \operatorname{Tr}(h(D + D')) + \operatorname{Tr}(G(D)D') - b\operatorname{Tr}(DD') + bN$$

with associated Fock matrix $F^b(D) = F(D) - bD$. The difference with the Roothaan algorithm is that for b large enough, there is a uniform gap $\gamma^b > 0$, and $D_k - D_{k+1}$ converges to 0 [5]. Therefore, we have the following theorems

Theorem 6.1 (Convergence of the Level-Shifting algorithm)**.** *Let $D_0 \in \mathcal{P}$. Then there exists $b_0 > 0$ such that for every $b > b_0$, the sequence D_k of iterates generated by the Level-Shifting algorithm with shift parameter b verifies the uniform well-posedness hypothesis with uniform gap $\gamma > 0$ and converges.*

Theorem 6.2 (Convergence rate of the Level-Shifting algorithm)**.** *Let D_k be the sequence of iterates generated by the Level-Shifting with shift parameter $b > b_0$, and let*

$$e_k = \sum_{l=k}^{\infty} \|D_{l+2} - D_l\|.$$

Then,

(1) If $\theta' = 1/2$ (non-degenerate case), then for any $\nu' < \frac{\gamma^b}{8\kappa'^2\|F^b\|_{op}^2}$, there exists $c > 0$ such that

$$e_k \leq c(1 - \nu')^k. \tag{6.1}$$

(2) If $\theta' \neq 1/2$ (degenerate case), then there exists $c > 0$ such that

$$e_k \leq ck^{-\frac{\theta'}{1-2\theta'}}. \tag{6.2}$$

We can use this result to heuristically predict the behavior of the algorithm when b is large. γ^b and $\|F^b\|_{op}$ both scale as b for large values of b. Assuming non-degeneracy, we can take $\kappa' > \frac{1}{\sqrt{2\lambda_1}}$, where λ_1 is the eigenvalue of smallest magnitude of the Hessian $H_1 + \frac{b}{2}H_2$, where $H_1 = H_{\mathcal{P}\times\mathcal{P}}E(D^\infty, D^\infty)$ and $H_2 = H_{\mathcal{P}\times\mathcal{P}}\|D - D'\|^2(D^\infty, D^\infty)$. But H_2 admits zero as an eigenvalue (for instance, note that $\|D - D'\|^2$ is constant along the curve $(D_t, D_t') = (U_t D U_t^T, U_t D' U_t^T)$, where U_t is a family of orthogonal matrices), so that, when b goes to infinity, λ_1 tends to the eigenvalue of smallest magnitude of H_1 restricted to the nullspace of H_2, and therefore stays bounded. Therefore, ν' scales as $\frac{1}{b}$, which suggests that b should not be too large for the algorithm to converge quickly.

7. Numerical results

We illustrate our results on atomic systems, using gaussian basis functions. The gradient method was implemented using the software Expokit [20] to compute matrix exponentials. In our computations, the cost of a gradient step is not much higher than a step of the Roothaan algorithm, since the limiting step is computing the Fock matrix, not the exponential. However, the situation may change if the Fock matrix is computed using linear scaling techniques. In this case, one can use more efficient ways of computing geodesics, as described in [8].

First, the Łojasiewicz inequality with exponent $\frac{1}{2}$ was checked to hold in the molecular systems and basis sets we encountered, suggesting that the minimizers are non-degenerate. Consequently, we never encountered sublinear convergence of any algorithm.

For a given molecular system and basis, we checked that the Level-Shifting algorithm converged as $(1 - \nu)^k$, where ν is asymptotically proportional to $\frac{1}{b}$, which we predicted theoretically in Section 6 (see Figure 7.1). This means that the estimates we used have at least the correct scaling behavior.
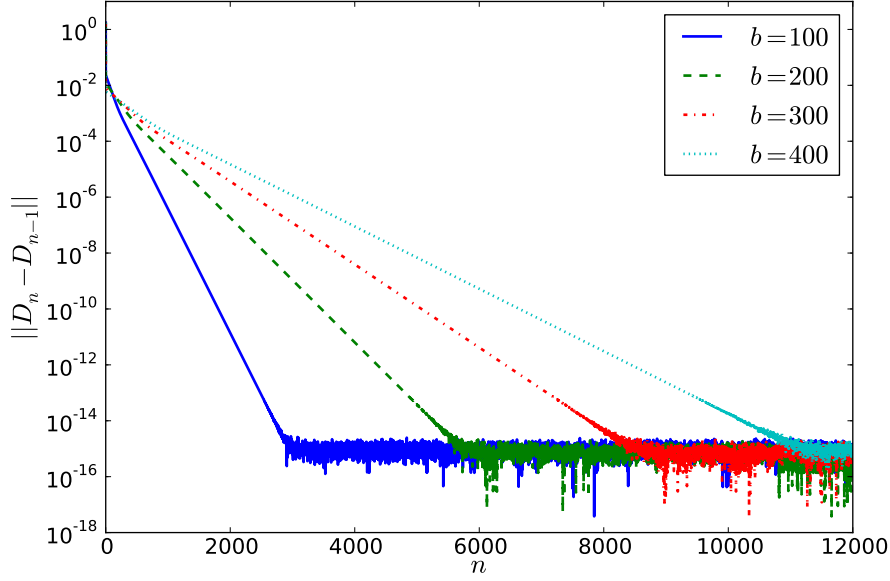
FIGURE 7.1. Convergence of the Level-Shifting algorithm. The convergence is linear until machine precision. The horizontal spacing of the curves reveals the asymptotic relationship $\nu \propto \frac{1}{b}$. The system considered is the carbon atom ($N = Z = 6$), under the RHF formalism, using the 3-21G gaussian basis functions.

Next, we compared the efficiency of the Roothaan algorithm and of the gradient algorithm, in the case where the Roothaan algorithm converges. Our analysis leads to the estimate $\nu = \frac{\gamma}{8\kappa'^2 \|F\|_{\mathrm{op}}^2}$ for the Roothaan algorithm, and $\nu = \frac{1}{4\kappa^2 \alpha}$ for the gradient algorithm with stepsize $t = \frac{1}{\alpha}$.

It is immediate to see that, up to a constant multiplicative factor, $\kappa' > \kappa$, $\gamma \leq \|F\|_{\mathrm{op}}$ and for the cases of interest $\alpha \approx \|F\|_{\mathrm{op}}$, so from our estimates we would expect the gradient algorithm to be faster than the Roothaan algorithm. However, in our tests the Roothaan algorithm was considerably faster than the gradient algorithm (see Figure 7.2). This conclusion holds even when the stepsize is adjusted at each iteration with a line search.

The reason that the Roothaan algorithm performs better than expected is that the inequality

$$\left\| [D_{k+2} - D_k, F(D_{k+1})] \right\| \leq 2 \left\| F(D_{k+1}) \right\| \left\| D_{k+2} - D_k \right\|$$

is very far from optimal. Whether an improved bound (in particular, one that does not depend on the dimension) can be derived is an interesting open question.

The outcome of these tests seems to be that the gradient algorithm is slower. It might prove to be faster in situations where the gap is small, or whenever κ' is much larger than κ. We have been unable to find concrete examples of such cases.

8. CONCLUSION, PERSPECTIVES

In this paper, we introduced an algorithm based on the idea of gradient descent. By using the analyticity of the objective function and of the constraint manifold, we were able to derive a Łojasiewicz inequality, and use that to prove the convergence of the gradient method, under the assumption of a small enough stepsize. Next, expanding on the analysis of [5], we extended the Łojasiewicz inequality to a Lyapunov function for the Roothaan algorithm. By linking the gradient of this Lyapunov function to the difference in the iterates of the algorithm, we proved convergence (or oscillation), an improvement over previous results which only prove a weaker version of this. In this framework, the Level-Shifting algorithm can be seen as a simple modification of the Roothaan algorithm, and as such can be treated by the same methods. In each case, we were also able to derive explicit bounds on the convergence rates.

The strength of the Łojasiewicz inequality is that no higher-order hypothesis are needed for its use. As a consequence, the rates of convergence we obtain weaken considerably if the algorithm converges to a degenerate critical point. A more precise study of the local structure of critical points is necessary to
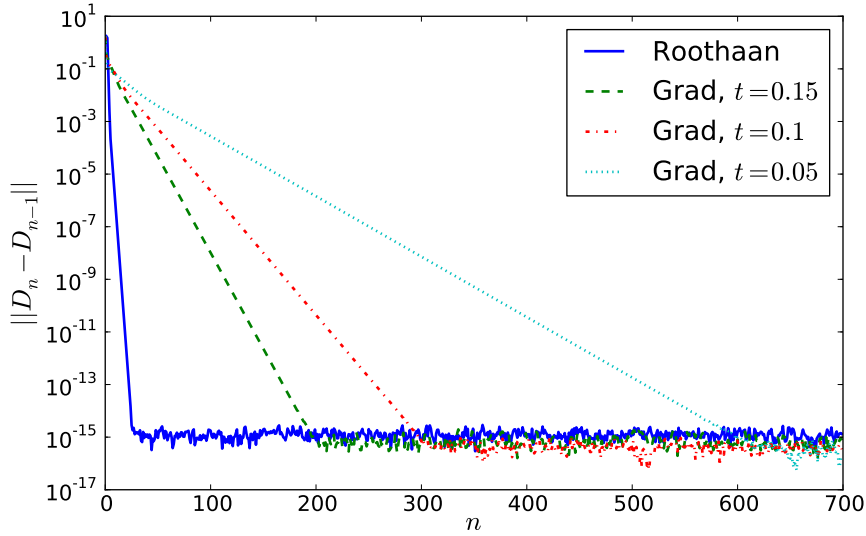
FIGURE 7.2. Comparison of Roothaan and gradient algorithm. The system considered is the carbon atom ($N = Z = 6$), under the RHF formalism, using the 3-21G gaussian basis functions.

understand why the algorithms usually exhibit geometric convergence. This is related to the problem of local uniqueness and is likely to be hard (and, indeed, to our knowledge has not been tackled yet).

Even though our results hide the complexity of the local structure in the constants of the Łojasiewicz inequality, they still provide insight as to the influence of the basis on the speed of convergence, and can be used to compare algorithms. All of our results use crucially the hypothesis of a finite-dimensional Galerkin space. For the gradient algorithm, we need it to ensure the existence of a stepsize that decreases the energy. This is analogous to a CFL condition for the discretization of the equation $\frac{\mathrm{d}D}{\mathrm{d}t} = -[D, [D, F_D]]$, and can only be lifted with a more implicit discretization of this equation. For the Roothaan and Level-Shifting algorithms, we use the finite dimension hypothesis to bound $F(D)$. As noted in Section 7, the inequality is not sharp, so it could be that the infinite-dimensional version of the Roothaan and Level-Shifting algorithms still converge. More research is needed to examine this.

The gradient algorithm we examined only converges towards a stationary point of the energy, that may not be a local minimizer, or even an Aufbau solution. However, it will generically converge towards a local minimizer, unlike the Level-Shifting algorithm with large b. Therefore, it is the most robust of the algorithms considered. Although it was found to be slower than other algorithms on the numerical tests we performed, it has the advantage that its convergence rate does not depend on the gap $\lambda_{N+1} - \lambda_N$, and might therefore prove useful in extreme situations.

An algorithm that could achieve the speed of the fixed-point algorithms with the robustness granted by the energy monotonicity seems to be the ODA algorithm of Cancès and Le Bris [5], along with variants such as EDIIS, or combinations of EDIIS and DIIS algorithms [12]. We were not able to examine these algorithms in this paper. At first glance, the ODA algorithm should fit into our framework (indeed, the ODA algorithm was built to satisfy an energy decrease inequality similar to (3.7)). However, it works in a relaxed parameter space $\widetilde{P}$, and using the commutator to control the differences of iterates as we did only makes sense on $\mathcal{P}$. Therefore, other arguments have to be used.

A variant on the gradient algorithm used here is to modify the local geometry of the manifold $\mathcal{P}$ by using a different inner product, leading to a variety of methods, including conjugate gradient algorithms [8]. These methods fit into our framework, as long as one can prove that they are "gradient-like", in the sense that one can control the gradient by the difference $D_{k+1} - D_k$. However, precise estimates of convergence rates might be hard to obtain.

Also missing from the present contribution is the study of other commonly used algorithms, such as DIIS [17], and variants of (quasi)-Newton algorithms [2, 11]. DIIS numerically exhibits a complicated behavior that is probably hard to explain analytically, and the (quasi)-Newton algorithms require a study of the second-order structure of the critical points, which we are unable to do.

12

9. Acknowledgments

References

[1] F. Alouges and C. Audouze. Preconditioned gradient flows for nonlinear eigenvalue problems and application to the Hartree-Fock functional. *Numerical Methods for Partial Differential Equations*, 25(2):380–400, 2009.

[2] G.B. Bacskay. A quadratically convergent Hartree–Fock (QC-SCF) method. Application to closed shell systems. *Chemical Physics*, 61(3):385–404, 1981.

[3] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: a primer. *Handbook of numerical analysis*, 10:3–270, 2003.

[4] E. Cancès and C. Le Bris. Can we outperform the DIIS approach for electronic structure calculations? *International Journal of Quantum Chemistry*, 79(2):82–90, 2000.

[5] E. Cancès and C. Le Bris. On the convergence of SCF algorithms for the Hartree-Fock equations. *Mathematical Modelling and Numerical Analysis*, 34(4):749–774, 2000.

[6] E. Cancès and K. Pernal. Projected gradient algorithms for Hartree-Fock and density matrix functional theory calculations. *The Journal of chemical physics*, 128:134108, 2008.

[7] E. Cancès. SCF algorithms for Hartree-Fock electronic calculations. In M. Defranceschi and C. Le Bris, editors, *Mathematical models and methods for ab initio quantum chemistry, Lecture Notes in Chemistry*, volume 74. Springer, 2000.

[8] A. Edelman, T.A. Arias, and S.T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20:303, 1998.

[9] J.B. Francisco, J.M. Martínez, and L. Martínez. Globally convergent trust-region methods for self-consistent field electronic structure calculations. *The Journal of chemical physics*, 121:10863, 2004.

[10] A. Haraux, M.A. Jendoubi, and O. Kavian. Rate of decay to equilibrium in some semilinear parabolic equations. *Journal of Evolution equations*, 3(3):463–484, 2003.

[11] S. Høst, J. Olsen, B. Jansík, L. Thøgersen, P. Jørgensen, and T. Helgaker. The augmented Roothaan–Hall method for optimizing Hartree–Fock and Kohn–Sham density matrices. *The Journal of chemical physics*, 129:124106, 2008.

[12] K.N. Kudin, G.E. Scuseria, and E. Cancès. A black-box self-consistent field convergence algorithm: One step closer. *The Journal of chemical physics*, 116:8255, 2002.

[13] E.H. Lieb and B. Simon. The Hartree-Fock theory for Coulomb systems. *Communications in Mathematical Physics*, 53(3):185–194, 1977.

[14] P.L. Lions. Solutions of Hartree-Fock equations for Coulomb systems. *Communications in Mathematical Physics*, 109(1):33–97, 1987.

[15] S. Łojasiewicz. *Ensembles semi-analytiques*. Institut des Hautes Etudes Scientifiques, 1965.

[16] R. McWeeny. The density matrix in self-consistent field theory. I. Iterative construction of the density matrix. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 235(1203):496, 1956.

[17] P. Pulay. Improved SCF convergence acceleration. *Journal of Computational Chemistry*, 3(4):556–560, 1982.

[18] J. Salomon. Convergence of the time-discretized monotonic schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(01):77–93, 2007.

[19] V.R. Saunders and I.H. Hillier. A "Level–Shifting" method for converging closed shell Hartree–Fock wave functions. *International Journal of Quantum Chemistry*, 7(4):699–705, 1973.

[20] R.B. Sidje. Expokit: a software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 24(1):130–156, 1998.

Université Paris-Dauphine, CEREMADE, Place du Maréchal Lattre de Tassigny, 75775 Paris Cedex 16, France.

E-mail address: `levitt@ceremade.dauphine.fr`